Nicolae Sfetcu

L’éthique des mégadonnées (Big Data) en recherche

Collection ESSAIS

MultiMedia Publishing

L'éthique des mégadonnées (Big Data) en recherche

Nicolae Sfetcu

16.03.2020

Sfetcu, Nicolae, « L'éthique des mégadonnées (Big Data) en recherche », SetThings (16 mars 2020), MultiMedia Publishing (ISBN : 978-606-033-348-7), DOI: 10.13140/RG.2.2.10128.56328, URL = https://www.telework.ro/fr/e-books/lethique-des-megadonnees-big-data-en-recherche/

Email: nicolae@sfetcu.com

Une traduction de :
Sfetcu, Nicolae, « Etica Big Data în cercetare », SetThings (6 iulie 2019), DOI: 10.13140/RG.2.2.27629.33761, MultiMedia Publishing (ed.), ISBN: 978-606-033-228-2, URL = https://www.telework.ro.com/ro/e-books/etica-big-data-in-cercetare/

Abstract

Les principaux problèmes rencontrés par les scientifiques qui travaillent avec des ensembles de données massives (mégadonnées, Big Data), en soulignant les principaux problèmes éthiques, tout en tenant compte de la législation de l'Union européenne. Après une brève *Introduction* au Big Data, la section *Technologie* présente les applications spécifiques de la recherche. Il suit une approche des principales questions philosophiques spécifiques dans *Aspects philosophiques,* et *Aspects juridiques* en soulignant les problèmes éthiques spécifiques du règlement de l'UE sur la protection des données 2016/679 (General Data Protection Regulation, « GDPR »). La section *Problèmes éthiques* détaille les problèmes spécifiques générés par le big data. Après une brève section de *Recherche de big data*, sont présentées les *Conclusions* sur l'éthique de la recherche dans l'utilisation du big data.

1. Introduction

Le terme *big data* désigne l'extraction, la manipulation et l'analyse des ensembles de données trop volumineux pour être traités de manière routinière. Pour cette raison, des logiciels spéciaux sont utilisés et, dans de nombreux cas, des ordinateurs et du matériel informatiques dédiés. Généralement, ces données sont analysées de manière statistique. Sur la base de l'analyse des données respectives, des prédictions de certains groupes de personnes ou d'autres entités sont généralement établies, en fonction de leur comportement dans diverses situations et à l'aide des techniques analytiques avancées. Ainsi, les tendances, les besoins et les évolutions comportementales de ces entités peuvent être identifiés. Les scientifiques utilisent ces données pour la recherche en météorologie, génomique, connectomique, (Nature 2008) simulations physiques complexes, biologie, protection de l'environnement, etc. (Reichman, Jones, and Schildhauer 2011)

Avec le volume croissant des données sur Internet, dans les médias sociaux, l'informatique en nuage, les appareils mobiles et les données gouvernementales, le big data constitue à la fois une menace et une opportunité pour les chercheurs de gérer et d'utiliser ces données, tout en maintenant les droits des personnes impliquées.

1.1 Définitions

Les mégadonnées comprennent généralement des ensembles de données dont les dimensions dépassent la capacité logicielle et matérielle habituelle, en utilisant des données non structurées, semi-structurées et structurées, l'accent étant mis sur les données non structurées. (Dedić and Stanier 2017) La taille du big data a augmenté depuis 2012, passant de dizaines de téraoctets à de nombreux exaoctets de données. (Everts 2016) Travailler efficacement avec le big

data implique l'apprentissage des machines pour détecter les modèles, (Mayer-Schönberger and Cukier 2014) mais ces données sont souvent un sous-produit d'autres activités numériques.

Selon une définition de 2018, « les mégadonnées sont des données qui nécessitent des outils informatiques parallèles pour gérer les données », ce qui constitue un tournant dans le domaine de l'informatique, qui repose sur les théories de la programmation parallèle et le manque de certitude des modèles précédents. Le big data utilise des statistiques inductives et concepts d'identification des systèmes non linéaires pour déduire des lois (régressions, relations non linéaires et effets causals) à partir de grands ensembles de données avec une faible densité d'informations afin d'obtenir des relations et des dépendances ou de prédire des résultats et des comportements.

Au niveau de l'Union européenne, il n'y a pas de définition obligatoire mais, selon l'avis 3/2013 du groupe de travail européen sur la protection des données,

> « Le terme « **mégadonnées** » se rapporte à l'augmentation considérable de l'accès aux informations et de leur utilisation automatisée: il fait référence aux volumes gigantesques de données numériques contrôlées par des entreprises, des gouvernements et d'autres grandes organisations, qui sont ensuite analysés de façon approfondie en utilisant des algorithmes. Les mégadonnées peuvent servir à établir des tendances générales et des corrélations, mais leur utilisation peut également toucher directement les individus ». (European Economic and Social Committee 2017)

Le problème avec cette définition est qu'elle n'envisage pas de réutiliser des données personnelles.

Règlement no. 2016/679 définit les **données à caractère personnel** (article 4, paragraphe 1) comme

> « toute information se rapportant à une personne physique identifiée ou identifiable (ci-après dénommée "personne concernée"); est réputée être une "personne physique identifiable" une personne physique qui peut être identifiée, directement ou indirectement, notamment par référence à un identifiant, tel qu'un nom, un numéro d'identification, des données de localisation, un identifiant en ligne, ou à un ou plusieurs éléments spécifiques propres à son identité physique, physiologique, génétique, psychique, économique, culturelle ou sociale. »

La définition s'applique également au niveau de l'UE aux personnes non identifiées, mais qui peut être identifiée en mettant en corrélation des données anonymes avec d'autres informations supplémentaires. Les données personnelles, une fois anonymisées (ou pseudo-anonymisées), peuvent être traitées sans autorisation préalable, en tenant compte toutefois du risque de réidentification de la personne concernée.

1.2 Les dimensions du big data

Les données sont partagées et stockées sur des serveurs, via l'interaction entre l'entité impliquée et le système de stockage. Dans ce contexte, les big data peuvent être classés en systèmes actifs (interaction synchrone, les données d'entité sont envoyées directement au système de stockage) et passifs (interaction asynchrone, les données sont collectées par un intermédiaire, puis introduits dans le système.

De plus, les données peuvent être transmises directement consciemment ou non (si la personne dont les données sont transmises n'est pas notifiée de manière claire et en temps voulu). Les données sont ensuite traitées pour générer des statistiques.

Selon la cible des analyses statistiques respectives, les dimensions des données peuvent être : a) individuelles (une seule entité est analysée); social (nous analysons des groupes d'entités distincts au sein d'une population; et hybrides (lorsqu'une entité est analysée en fonction de son appartenance à un groupe déjà défini).

L'énorme production actuelle de données générées par les utilisateurs devrait augmenter de 2000% dans le monde d'ici 2020 et est souvent non structurée. (Cumbley and Church 2013) En général, le big data est caractérisé par:

- Volume (quantité de données) ;
- Variété (produits de différentes sources dans différents formats) ;

- Vitesse (rapidité d'analyse des données en ligne) ;
- Véracité (les données sont incertaines et doivent être vérifiées) ;
- Valeur (évaluée par analyse).

Le volume de données produites et stockées évolue actuellement de manière exponentielle, plus de 90% d’entre elles ayant été générées au cours des quatre dernières années. (European Economic and Social Committee 2017) Les volumes importants nécessitent une analyse rapide, avec un fort impact sur la véracité. Des données incorrectes peuvent causer des problèmes lorsqu'elles sont utilisées dans le processus de décision.

Un des problèmes majeurs du big data est de savoir si des données complètes sont nécessaires pour tirer certaines conclusions sur leurs propriétés, ou si un échantillon est suffisant. Big data contient dans son nom un terme lié à la taille, qui est une caractéristique importante du big data. Cependant, l'échantillonnage (statistique) permet de sélectionner des points de collecte de données corrects parmi un ensemble plus large afin d'estimer les caractéristiques de la population entière. Le big data peut être échantillonné à travers différentes catégories de données lors du processus de sélection de l'échantillon à l'aide d'algorithmes d'échantillonnage pour le big data.

2. La technologie

Les données doivent être traitées avec des outils de collecte et d'analyse avancés, basés sur des algorithmes prédéterminés, afin d'obtenir des informations pertinentes. Les algorithmes doivent également prendre en compte les aspects invisibles pour les perceptions directes.

En 2004, Google a publié un article sur un processus appelé MapReduce, qui propose un modèle de traitement parallèle. (Dean and Ghemawat 2004) MIKE 2.0 est également une application open source pour la gestion de l’information. (MIKE2.0 2019) Plusieurs études de

2012 ont montré que l'architecture optimale pour traiter les problèmes liés au big data repose sur plusieurs couches. Une architecture parallèle distribuée distribue les données sur plusieurs serveurs (environnements d'exécution parallèle), ce qui améliore considérablement les vitesses de traitement des données.

Selon un rapport publié par le McKinsey Global Institute en 2011, les principaux composants et écosystèmes du big data sont : (Manyika et al. 2011) des techniques d'analyse des données (apprentissage automatique, traitement du langage naturel, etc.), des grandes technologies (informatique décisionnelle, informatique en nuage, bases de données) et vues (charts, graphiques, autres vues de données).

Les mégadonnées fournissent des informations en temps réel ou quasi réel, évitant ainsi le temps de latence autant que possible.

2.1 Applications

Les mégadonnées dans les processus gouvernementaux augmentent la rentabilité, la productivité et l'innovation. Les registres d'état civil sont une source pour le big data. Les données traitées aident dans des domaines critiques du développement, tels que les soins de santé, l'emploi, la productivité économique, la criminalité, la sécurité et la gestion des catastrophes naturelles et des ressources. (Kvochko 2012)

Big data fournit également une infrastructure permettant de mettre en évidence les incertitudes, les performances et la disponibilité des composants. Les tendances et les prévisions dans l'industrie nécessitent une grande quantité de données et des outils de prévision avancés.

Le big data contribue à l'amélioration des soins de santé en fournissant des médicaments personnalisés et des analyses prescriptives, des interventions cliniques avec évaluation du risque et analyse prédictive, etc. Le niveau de données générées dans les systèmes de santé est très élevé.

Cependant, la génération de « données altérées » pose un problème urgent, qui augmente avec le volume de données, d'autant plus que la plupart d'entre elles sont non structurées et difficiles à utiliser. L'utilisation du big data dans les soins de santé a généré d'importants défis éthiques, ayant des implications pour les droits individuels, la vie privée et l'autonomie, la transparence et la confiance.

Dans les médias et la publicité, pour le big data sont utilisés de nombreux points d'information sur des millions de personnes pour servir ou transmettre des messages ou des contenus personnalisés.

Dans le domaine de l'assurance maladie, des données sont collectées sur les « déterminants de la santé », ce qui aide à élaborer des prévisions sur les coûts de la santé et à identifier les problèmes de santé des clients. Cette utilisation est controversée, en raison de la discrimination des clients ayant des problèmes de santé. (Allen 2018)

Les mégadonnées et les technologies de l'information se complètent, aidant ensemble à développer l'Internet des objets (IoT) pour interconnecter des appareils intelligents et collecter des données sensorielles utilisées dans différents domaines.

En sport, le big data peut aider à améliorer l'entraînement et la compréhension des compétiteurs à l'aide de capteurs spécifiques et à prédire les performances futures des athlètes. Les capteurs attachés aux voitures de Formule 1 collectent, entre autres choses, les données de pression des pneus pour rendre la consommation de carburant plus efficace.

2.1.1 En recherche

En science, les systèmes big data sont utilisés de manière intensive dans les accélérateurs de particules du CERN (150 millions de capteurs transmettent des données 40 millions fois par seconde, pour environ 600 millions de collisions par seconde, dont ils ne sont utilisés qu'après

filtrage que 0,001% du total des données obtenues), (Brumfiel 2011) dans des radiotélescopes astrophysiques construits à partir de milliers d'antennes, décodant le génome humain (au départ, cela prenait quelques années, avec le big data peut être réalisé en moins d'une journée), des études climatiques, etc.

Les grandes entreprises informatiques utilisent des entrepôts de données de l'ordre de plusieurs dizaines de pétaoctets pour la recherche, les recommandations et le marchandisage. La plupart des données sont collectées par Facebook, avec plus de 2 milliards d'utilisateurs actifs mensuels, (Constine 2017) et Google, avec plus de 100 milliards de recherches par mois. (Sullivan 2015)

Dans la recherche on utilise beaucoup l'interrogation cryptée et la formation de grappes dans le big data. Les pays développés investissent actuellement beaucoup dans la recherche sur le big data. Au sein de l'Union européenne, ces recherches sont incluses dans le programme Horizon 2020. (European Commission 2019)

Les programmes de recherche utilisent souvent des ressources API de Google et Twitter pour accéder à leurs systèmes big data, gratuitement ou avec paiement.

Les grands ensembles de données comportent des problèmes d'algorithme qui n'existaient pas auparavant et il est impératif de changer fondamentalement les méthodes de traitement. À cette fin, des ateliers spéciaux ont été créés. Ils réunissent des scientifiques, des statisticiens, des mathématiciens et des praticiens pour débattre les défis algorithmiques du big data.

3. Aspects philosophiques

Le big data peut générer, par inférences, de nouvelles connaissances et perspectives. Le paradigme qui résulte de l'utilisation du big data crée de nouvelles opportunités.

L'une des principales préoccupations dans le cas du big data est que les scientifiques des données ont tendance à travailler avec des données sur des sujets qu'ils ne connaissent pas et n'ont jamais été en contact, étant éloignés du produit final de leur activité (l'application des analyses). Une étude récente (Tanner 2014) indique que cela peut être la raison d'un phénomène connu sous le nom d'aliénation numérique.

Le big data a une grande influence au niveau gouvernemental, affectant positivement la société. Ces systèmes peuvent être rendus plus efficaces en appliquant des politiques de transparence et de gouvernance ouverte, telles que les données ouvertes.

Après avoir développé des modèles prédictifs pour le comportement du public cible, le big data peut être utilisé pour générer des alertes précoces pour diverses situations. Il y a donc un retour positif entre la recherche et la pratique, avec des découvertes rapides tirées de la pratique.

A. Richterich déclare que la popularité de la surveillance de l'activité des utilisateurs a été motivée par des affirmations selon lesquelles l'utilisation (et la collecte de données avec) ces appareils amélioreraient le bien-être, la santé et l'espérance de vie des utilisateurs, et réduiraient considérablement les coûts des soins de santé. (Richterich, 2018) Pour obtenir le consentement de l'utilisateur, de nombreuses entreprises ont offert des remises aux clients qui seraient disposés à donner accès à leurs données de surveillance. (Mearian 2015) Mais il y a aussi des préoccupations concernant l'influence de ces technologies sur la société, en particulier dans les questions liées à l'équité, la discrimination, la vie privée, l'abus de données et la sécurité. (Collins 2016)

Conceptuellement, le big data doit être compris comme un terme générique désignant un ensemble de technologies émergentes. Dans leur utilisation, nous devons prendre en compte les contextes culturels, sociaux et technologiques, les réseaux, les infrastructures et les interdépendances qui peuvent avoir un sens sur le big data. Le terme « big data » fait référence

non seulement aux données en tant que telles, mais également aux pratiques, infrastructures, réseaux et politiques qui influencent leurs diverses manifestations. Comprendre les mégadonnées comme un ensemble de technologies émergentes semble être conceptuellement utile, car il comprend des développements numériques activés dans la collecte, l'analyse et l'utilisation des données. (Richterich, 2018)

Dans ce contexte, Rip décrit le dilemme des développements technologiques : "Pour les technologies émergentes avec leur avenir indéterminé, il y a le défi d'articuler des valeurs et des règles appropriées qui auront du poids. Cela se produit en articulant des promesses et des visions sur les nouvelles technologies." (Rip 2013, 192) Ainsi, les technologies émergentes sont des lieux de « normativité omniprésente » caractérisés par l'articulation de promesses et de peurs, conceptualisant une telle « normativité omniprésente » comme une approche « dans l'esprit d'une éthique pragmatique, dans laquelle les positions normatives co-évoluent. » (Rip 2013)

L'éthique pragmatique souligne que les nouvelles technologies se développent dans des sociétés où elles sont associées / dissociées de manière discursive par certaines normes et valeurs. Dans le même temps, le pragmatisme affirme que l'augmentation du grand nombre de données et de pratiques liées à la recherche n'est pas une simple question de supériorité technologique. Ils forment un champ de justification normative et de contestation.

Les néo-pragmatiques dans l'approche de l'éthique aborde les connaissances épistémologiques à travers la falsification des connaissances (scientifiques), avec des évaluations critiques des structures du pouvoir social. Keulartz et al. ont proposé une approche pragmatique de l'éthique dans une culture technologique (Keulartz et al. 2004) "comme une alternative qui combine les forces de l'éthique appliquée et des études scientifiques et technologiques, tout en évitant les lacunes de ces domaines." (Richterich, 2018) Ainsi, l'éthique appliquée est une approche

efficace en termes de détection et d'expression des normes impliquées dans des actions (inter) sociotechniques ou résultant d'actions sociotechniques, mais elle n'a aucune possibilité de capter la normativité inhérente et l'agent des technologies. (Keulartz et al. 2004, 5)

Keulartz et al. considère que le manque d'évaluations technologiques normatives peut ainsi être surmonté: "l'impasse qui a émergé de ce point de vue" (c'est-à-dire les "angles morts" de l'éthique appliquée) peut être surmontée par une réévaluation du pragmatisme. (Keulartz et al. 2004, 14) Le pragmatisme éthique peut être caractérisé par trois principes communs: l'anti-fondationalisme, l'anti-dualisme et l'anti-scepticisme.

L'anti-fondationalisme fait référence au principe de falsifiabilité, étant donné que nous ne pouvons pas atteindre la certitude en termes de connaissances ou de valeurs ("vérité finale"), mais les connaissances, ainsi que les valeurs et les normes, changent avec le temps. Les valeurs morales ne sont pas statiques, mais peuvent être renégociées en fonction des évolutions technologiques.

L'anti-dualisme implique la nécessité de s'abstenir de dichotomies prédéterminées. Parmi les dualismes critiqués par Keulartz figurent l'essence / l'apparence, la théorie / la pratique, la conscience / la réalité et les faits / la valeur. L'éthique appliquée a tendance à assumer de tels dualismes *a priori*, par opposition au pragmatisme, qui souligne les interrelations et les frontières floues entre ces catégories.

L'anti-scepticisme est étroitement lié à la nécessité des perspectives situées et d'une normativité explicite, liées au fondement anti-cartésien du pragmatisme.

Dans la recherche européenne, le pragmatisme était généralement rejeté comme "superficiel et opportuniste", étant associé à des "stéréotypes négatifs", (Joas 1993) étant accusé "d'utilitarisme et de méliorisme". (Keulartz et al. 2004, 15) À la fin des années 1990 et 2000, le pragmatisme a connu un renouveau dans la recherche européenne. (Baert and Turner 2004)

L'analyse du big data d'un point de vue éthique implique deux principaux aspects interdépendants : un aspect théorique (la description philosophique des éléments soumis au contrôle éthique) et une vision pragmatique (de l'impact sur la vie des personnes et des organisations). (European Economic and Social Committee 2017)

Il y a des problèmes éthiques causés par l'intelligence artificielle, et un lien étroit entre le big data et l'intelligence artificielle et ses dérivés : apprentissage automatique, analyse sémantique, exploitation des données.

Une approche éthique passe par l'agence morale avec au moins les trois conditions de causalité, de connaissance et de choix. Selon Noorman: (Noorman 2012)

- Il existe des liens de causalité entre les personnes et le résultat des actions. La responsabilité de la personne découle du contrôle du résultat.
- Le sujet doit être informé, y compris les conséquences possibles.
- Le sujet doit donner son consentement et agir d'une certaine manière.

Le professeur Floridi, dans *La quatrième révolution*, identifie le problème moral du big data avec la découverte d'un modèle simple : une nouvelle frontière d'innovation et de concurrence. (Floridi 2014) Un autre problème associé au big data est le risque de découvrir ces modèles, changeant ainsi les prédictions.

La règle de base de l'éthique du big data est la protection de la vie privée, la liberté et la discrétion de décider de manière autonome. Il convient de noter qu'il existe une tension continue entre les besoins individuels et ceux d'une communauté.

Il est possible d'identifier plusieurs problèmes éthiques liés à l'exploitation des big data (European Economic and Social Committee 2017),

- *Confidentialité* - La limite extrême de la confidentialité est l'isolement, défini par Alan F. Westin comme "le retrait volontaire d'une personne de la société en général par des moyens physiques dans un état de solitude". Moor et Tavani ont défini un modèle de confidentialité appelé. Contrôle restreint de l'accès (RALC) qui différencie la confidentialité, la justification et la gestion de la confidentialité.
- *Réalité adaptée et bulles de filtrage* - L'application sur un serveur collecte des informations en tirant des enseignements, puis utilise ces informations pour construire un modèle de nos intérêts. Lorsqu'un système utilise ces modèles pour filtrer des informations, nous pouvons être amenés à croire que ce que nous voyons est une vue complète d'un contexte spécifique, alors qu'en fait nous sommes limités par la "compréhension" d'un algorithme qui a construit le modèle. Les effets éthiques peuvent être multiples : certaines informations peuvent être cachées, imposer des préjugés que nous ne connaissons pas, notre vision du monde peut devenir progressivement limitée, et à long terme elle pourrait générer un certain point de vue.
- *Gestion des données après le décès* - Qu'advient-il des données d'un utilisateur décédé ? Les héritiers deviennent-ils leurs propriétaires ? Les données peuvent-elles être retirées du monde numérique ? Il y a ici des problèmes juridiques et technologiques.
- *Biais d'algorithme* - L'interprétation des données implique presque toujours certains biais. De plus, il est possible qu'une erreur dans un algorithme introduise des formes de biais. Un problème éthique est notre confiance implicite dans les algorithmes, avec des risques élevés lorsque les risques ne sont pas pris en compte en raison d'erreurs de programmation ou d'exécution des algorithmes.

- *Confidentialité vs accroître la puissance de l'analyse* - Elle se réfère à la nature émergente de l'information comme un système complexe : le résultat de données provenant de différents contextes est plus que la simple somme des parties.
- *Limitation de la finalité* - Il est très difficile, voire impossible, de limiter l'utilisation des données. La confidentialité n'est pas un bloc unique, il y a des formes subtiles de perte de la confidentialité.
- *Inertie du profil numérique des utilisateurs* - Il s'agit du sujet de la réalité personnalisée. Un modèle qui implique les intérêts d'un utilisateur est généralement basé sur le comportement passé et les informations passées. Ainsi, les algorithmes ne sont pas basés sur l'identité réelle de la personne, mais sur une version antérieure. Cela influencera le comportement réel de l'utilisateur, poussé à maintenir ses anciens intérêts et donc à ne pas pouvoir découvrir d'autres opportunités. Si l'utilisateur n'est pas conscient de ce problème, l'influence de l'inertie sera beaucoup plus importante.
- *Radicalisation des utilisateurs, conformisme et sectarisme* - les big data peuvent se forger des opinions en utilisant des algorithmes de filtrage / recommandation, des informations, des articles et des messages personnalisés et des recommandations spécifiques d'amis. Ainsi, les utilisateurs seront de plus en plus en contact avec les personnes, les opinions et les faits qui soutiendront leur position d'origine. Cette tendance est souvent cachée aux utilisateurs de systèmes basés sur le big data, avec la tendance à développer des préjugés, allant de la conformité à la radicalisation. Il est possible de postuler la formation d'une sorte de subconscient technologique ayant un impact sur le développement de la personnalité des utilisateurs, phénomène évident dans le cas des réseaux sociaux, où la distance entre le monde réel ("physique") et Internet est fortement atténuée.

- *Impact sur les capacités personnelles et la liberté*
- *Égalité des droits entre le propriétaire des données et l'opérateur de données* - Habituellement, la personne dont les données sont utilisées n'est pas leur propriétaire légal. Par conséquent, une condition minimale est que cette personne ait accès à ses propres données, lui permettant de les télécharger et éventuellement de les supprimer.

4 Aspects juridiques

L'utilisation des mégadonnée présente des problèmes juridiques importants, notamment en termes de protection des données. Le cadre juridique existant de l'Union européenne, basé notamment sur la Directive 46/95/CE et le Règlement général sur la protection des données (RGPD - en anglais : General Data Protection Regulation, GDPR) assurent une protection adéquate. Mais pour les mégadonnées, une stratégie globale et complète est nécessaire. L'évolution au fil du temps est passée du droit d'exclure les autres au droit de contrôler leurs propres données et, à l'heure actuelle, à repenser le droit à l'identité (numérique).

La collecte et l'agrégation des données dans les mégadonnées ne sont pas soumises à la réglementation sur la protection des données, en raison des nouvelles perspectives sur la confidentialité, avec la possibilité de formes spécifiques de discrimination.

En 2014, le rapport de Podesta concluait que

> « L’analyse des mégadonnées peut potentiellement occulter les mécanismes de protection des droits civiques qui existent depuis longtemps dans la façon d’utiliser les informations à caractère personnel en matière de logement, de crédit, d’emploi, de santé, d’éducation, et sur le marché ». (European Economic and Social Committee 2017)

Il s'ensuit que de nouveaux moyens spécifiques de protéger les citoyens sont nécessaires, car le cadre juridique, même s'il est théoriquement applicable, ne semble pas offrir une protection adéquate et complète.

4.1 RGPD (GDPR)

Le règlement de l'UE sur la protection des données 2016/679 (Règlement général sur la protection des données, « RGPD ») traite la protection des données et la vie privée des personnes dans l'Union européenne et l'Espace économique européen. Il traite spécifiquement l'exportation de données personnelles en dehors des zones de l'UE et de l'EEE. Le RGPD entend simplifier l'environnement réglementaire en unifiant la réglementation au sein de l'UE. (European Parliament 2016)

Le RGPD s'applique dans deux cas au traitement des données personnelles (a) l'accès aux biens ou services aux frais des personnes dans l'UE, ou (b) le suivi de leur comportement au sein de l'UE. Ainsi, le règlement permet de l'étendre à tous les fournisseurs de services Internet, même s'ils ne sont pas établis dans l'UE. Plus généralement, le RGPD s'applique à tous les grands agrégateurs de données, quelles que soient les connexions géographiques ou physiques.

Étapes du traitement des données personnelles

Le traitement des données personnelles est défini par l'article 4, paragraphe 2, comme « toute opération ou tout ensemble d’opérations effectuées ou non à l’aide de procédés automatisés et appliquées à des données ou des ensembles de données à caractère personnel, telles que la collecte, l’enregistrement, l’organisation, la structuration, la conservation, l’adaptation ou la modification, l’extraction, la consultation, l’utilisation, la communication par transmission, la diffusion ou toute autre forme de mise à disposition, le rapprochement ou l’interconnexion, la limitation, l’effacement ou la destruction. »

Les mégadonnées comprennent plusieurs activités de traitement des données personnelles, chacune avec ses propres règles spécifiques :

1. Collecte de données

2. Stockage des données
3. Agrégation de données
4. Analyse des données et utilisation des résultats d'analyse

Principes du traitement des données

Le traitement des données est basé sur les principes suivants énoncés à l'article 5 du RGPD :

1. Légalité, équité et transparence : les utilisateurs doivent être pleinement et correctement informés de la politique de confidentialité et pouvoir accéder facilement à leurs propres données.
2. Limitation de la finalité : les collecteurs de données doivent informer la personne concernée des finalités de la collecte de données, qui peuvent être traitées ultérieurement à ces seules fins.
3. Minimisation des données : seules les données personnelles pertinentes aux fins déclarées seront collectées.
4. Exactitude et mise à jour : Les données seront mises à jour et rectifiées chaque fois que cela est requis par l'objectif déclaré. Dans le cas des mégadonnées, le droit des utilisateurs d'annuler ou de supprimer des données personnelles est très important.
5. Limitation du stockage : les données ne seront stockées que pendant le traitement et ensuite détruites. La durée de stockage peut être prolongée dans la mesure où les données sont archivées pour l'intérêt public, la recherche scientifique ou historique ou à des fins statistiques.
6. Intégrité et confidentialité : l'opérateur des données garantit une sécurité adéquate des données personnelles grâce à des mesures techniques et organisationnelles.

Politique de confidentialité et transparence

Dans le cas de la collecte de données afin de remplir un formulaire, le principe de minimisation des données sera respecté, seules les données pertinentes et strictement nécessaires étant demandées. Dans le cas d'une collecte automatique de données, comme les cookies, la surveillance Web ou la géolocalisation, la politique de confidentialité doit informer l'utilisateur de cet aspect.

Finalités du traitement des données

Des données anonymes et agrégées peuvent être traitées pour identifier le comportement de certaines catégories de consommateurs. À cette fin, l'opérateur de données procède à l'anonymisation puis les transfère à un tiers les utilisant.

Confidentialité par conception et confidentialité implicite

Les concepts de confidentialité par conception et confidentialité implicite n'étaient pas explicitement inclus dans les réglementations de l'UE. Mais, selon l'art. 78 du RGPD,

> « [a]fin d’être en mesure de démontrer qu’il respecte le présent règlement, le responsable du traitement devrait adopter des règles internes et mettre en œuvre des mesures qui respectent, en particulier, les principes de protection des données dès la conception et de protection des données par défaut. Ces mesures pourraient consister, entre autres, à réduire à un minimum le traitement des données à caractère personnel, à pseudo-anonymiser les données à caractère personnel dès que possible, à garantir la transparence en ce qui concerne les fonctions et le traitement des données à caractère personnel, à permettre à la personne concernée de contrôler le traitement des données, à permettre au responsable du traitement de mettre en place des dispositifs de sécurité ou de les améliorer. Lors de l’élaboration, de la conception, de la sélection et de l’utilisation d’applications, de services et de produits qui reposent sur le traitement de données à caractère personnel ou traitent des données à caractère personnel pour remplir leurs fonctions, il convient d’inciter les fabricants de produits, les prestataires de services et les producteurs d’applications à prendre en compte le droit à la protection des données lors de l’élaboration et de la conception de tels produits, services et applications et, compte dûment tenu de l’état des connaissances, à s’assurer que les responsables du traitement et les sous-traitants sont en mesure de s’acquitter des obligations qui leur incombent en matière de protection des données. » (European Parliament 2016)

Le paradoxe (juridique) des mégadonnées

L'utilisation des mégadonnées implique au moins un paradoxe : d'une part, les mégadonnées garantit une transparence maximale mais en même temps, il n'y a pas de transparence adéquate concernant l'utilisation des mégadonnées. La transparence est une question fondamentale car elle influence la capacité d'un utilisateur à autoriser la divulgation de ses informations. (European Economic and Social Committee 2017)

5. Problèmes éthiques

L'éthique des mégadonnées implique l'adhésion aux concepts de bons et mauvais comportements concernant les données, en particulier les données personnelles. L'éthique des mégadonnées se concentre sur les collecteurs et diffuseurs des données structurés ou non structurés.

L'éthique des mégadonnées est soutenue, au niveau de l'UE, par une documentation complète, qui cherche à trouver des solutions concrètes pour maximiser la valeur des mégadonnées sans sacrifier les droits humains fondamentaux. Le Contrôleur européen de la protection des données (CEPD) soutient le droit à la vie privée et le droit à la protection des données personnelles dans le respect de la dignité humaine. Selon ces documents, le conflit conceptuel entre confidentialité et mégadonnées, et entre confidentialité et innovation, doit être surmonté. Il est essentiel d'identifier les moyens d'intégrer la dimension éthique dans la conception des innovations. (European Economic and Social Committee 2017)

Selon le nouveau règlement de l'UE 2016/679, les opérateurs de données doivent mettre en œuvre les mesures et technologies de confidentialité pour améliorer la confidentialité lors de la détermination des modalités de traitement et du traitement lui-même. Grâce à l'Agence de l'Union européenne chargée de la sécurité des réseaux et de l'information (ENISA), de nombreuses

stratégies de confidentialité ont été identifiées par leur conception (minimisation des données, masquage des données personnelles et leurs interconnexions, traitement séparé des données personnelles, choix du plus haut niveau d'agrégation, transparence, surveillance, politique de confidentialité, problèmes juridiques).

Un moyen de base pour une coexistence pacifique entre l'exploitation des mégadonnées et la protection des données est le contrôle par l'utilisateur des données personnelles, ce qui conduit à la transparence et à la confiance entre les utilisateurs et les fournisseurs de services numériques. Comme indiqué dans l'analyse d'impact du RGPD,

> « L’instauration d’un climat de confiance dans l’environnement en ligne est essentielle au développement économique. En effet, s’ils n’ont pas totalement confiance, les consommateurs hésiteront à faire des achats en ligne et à recourir à de nouveaux services, y compris aux services administratifs en ligne. Si ce manque de confiance n’est pas résolu, il continuera de ralentir l’innovation dans l’utilisation des nouvelles technologies, d’entraver la croissance économique et de priver le secteur public des avantages potentiels de la numérisation de ses services. »

Dans le cas des mégadonnées, les modèles de consentement traditionnels sont insuffisants et obsolètes. « Le consentement doit être assez détaillé pour couvrir l’ensemble des traitements et des finalités du traitement et la réutilisation des données à caractère personnel. » (European Economic and Social Committee 2017)

Un problème particulier est la *portabilité* des données, soutenue au niveau de l'UE par le CEPD dans l'avis 7/2015, (MORO 2016) où il est nécessaire de garantir le droit des citoyens d'accéder et de corriger les données personnelles grâce à un contrôle étendu. La portabilité des données peut aider à accroître la sensibilisation et le contrôle des consommateurs en transférant des services en ligne.

CEPD estime que les données à caractère personnel devraient être traitées ainsi que d'autres ressources importantes, telles que le pétrole, où les échanges ont lieu entre les parties également informés (symétrie de l'information). En fait, le marché des informations personnelles a un

caractère d'asymétrie d'information, n'étant ni transparent ni équitable, les clients ne sont pas rémunérés pour les informations personnelles qu'ils fournissent. Ainsi, la portabilité des données favoriserait un environnement plus compétitif entre les bénéficiaires de ces données, les utilisateurs pouvant choisir à qui ils donnent les données personnelles.

Une autre approche implique le *magasin* des données personnelles, avec la possibilité pour l'utilisateur d'accorder ou de retirer le consentement pour ses données personnelles. (MORO 2016) (DG Connect 2015) Le magasin des données personnelles implique un « concept, un cadre et une architecture de mise en œuvre qui fait passer l’acquisition et le contrôle des données d’un modèle de données distribué à un modèle centré sur l’utilisateur. » (European Economic and Social Committee 2017) La portabilité des données pourrait garantir cela.

Le CEPD soutient la promotion des bénéficiaires responsables et la réduction de la bureaucratie dans la protection des données, grâce à des codes de conduite, des audits, des certifications et une nouvelle génération de clauses contractuelles et de règles d'entreprise obligatoires. La responsabilité des bénéficiaires des mégadonnées implique la mise en place de politiques internes et de systèmes de contrôle conformes à la législation en vigueur, à travers des solutions intelligentes et dynamiques qui garantissent le respect des principes fondamentaux (minimisation des données, limitation des finalités, qualité des données, traitement correct et transparent des données, conception, limitation du stockage, intégrité et confidentialité).

L'éthique des données est basée sur les principes suivants: *propriété* (les individus sont propriétaires de leurs données), *transparence des transactions* (les utilisateurs doivent avoir un accès transparent à la conception de l'algorithme), *consentement* (l'utilisateur doit être informé et consentir expressément à l'utilisation des données personnelles), *confidentialité* (la confidentialité de l'utilisateur doit être protégée), *financière* (l'utilisateur doit connaître les transactions financières

résultant de l'utilisation de ses données personnelles) et l'*ouverture* (les ensembles de données agrégées doivent être librement disponibles).

L'éthique dans la recherche

Le terme *étude des données critiques* implique que les chercheurs étudient les mégadonnées d'un point de vue critique. L'étude des données dans ce contexte implique, en plus de leur analyse, l'intégration des données dans les pratiques (connaissances), les institutions et systèmes politiques et économiques, à travers l'interaction complexe entre les données et les entités qui les produisent, les détiennent et les utilisent.

Un rapport de l'OCDE (2013) souligne que, contrairement aux normes éthiques appliquées aux données de recherche communes, dans le cas des mégadonnées : (OECD 2013)

- La collecte des données n'a pas fait l'objet d'un processus officiel d'examen éthique.
- Les règles éthiques communes ne seront pas mises en œuvre dans le cas des mégadonnées.
- L'utilisation des données de recherche peut différer de l'objectif initial.
- Les données ne sont plus conservées sous forme d'ensembles discrets.

La relation entre ceux qui fournissent les données et ceux qui les utilisent est souvent indirecte et variable. Un rapport plus récent de l'OCDE (2016) soutient que cette relation est plus faible ou inexistante, les mégadonnées limitant les capacités habituelles. (OECD 2016)

Le magasin des données est important pour l'intégrité de la recherche. Les données doivent avoir une « provenance » claire, avec des sources et un traitement connu et identifié.

De nombreuses données qui ne sont pas spécifiquement collectées pour la recherche ont des normes différentes dans la recherche de données.

Pour certaines données, souvent avec de valeur commerciale (par exemple, les données collectées sur Twitter), leur reproduction est soumise à des restrictions légales. (UK Data Service 2017)

Les magasins de données doivent respecter des normes de transparence et de reproductibilité.

Prise de conscience

La prise de conscience du type de données fournies lors d'une inscription en ligne (pour la création d'un compte ou d'un abonnement, par exemple) est un fait rare, d'autant plus qu'il existe la possibilité d'utiliser une identité numérique existante (profil Facebook par exemple) au lieu d'une liste séparée pour un accès plus rapide. De telles situations créent une opacité concernant les données partagées entre le fournisseur d'identité et le service utilisé. (European Economic and Social Committee 2017)

Consentement

Afin d'utiliser les données personnelles d'une personne, son consentement informé et explicite est requis concernant qui, quand, comment et dans quel but elles sont utilisées. Lorsque les données doivent être partagées, ces utilisations doivent être portées à la connaissance de la personne. Il devrait toujours être possible de retirer son consentement pour une utilisation future.

Dans l'analyse des mégadonnées, très peu de choses peuvent être connues sur les utilisations futures prévues des données, ainsi que sur les avantages et les risques impliqués. Ici, il existe des procédures de consentement « large » et « générique » pour partager des données génomiques, par exemple, et à des fins différentes. Même lorsqu'il est fait correctement, il existe des défis pratiques spécifiques : obtenir un consentement éclairé peut être impossible ou très

coûteux, et la validité du consentement est contestée lorsque l'accord est nécessaire pour accéder à un service.

Contrôle

Dans le monde d'aujourd'hui, les données personnelles peuvent être échangées comme n'importe quelle devise dans la mise en œuvre des mégadonnées. Il y a différentes opinions sur la mesure dans laquelle cette situation est éthique, y compris sur qui participer au profit obtenu de ces transactions.

Dans le modèle d'échange des données personnelles, la transmission des données personnelles est un cadre qui donne aux gens la possibilité de contrôler leur identité numérique et de créer des accords de partage des données granulaires.

L'idée des données ouvertes, centrée sur l'argument selon lequel les données devraient être disponibles gratuitement, est en train d'émerger. La volonté de partager des données varie selon la personne.

Dans le cas des enfants, les parents ou tuteurs ont la responsabilité de leurs données, qui ne peuvent pas être échangées contre des avantages financiers.

Au niveau national, un gouvernement est souverain sur les données générées et collectées. Le 26 octobre 2001, la Loi patriotique est entrée en vigueur aux États-Unis, et le 25 mai 2018, le Règlement Général de la Protection des Données 2016/679 (RGPD) au niveau de l'Union européenne, pour les questions liées à la protection des données personnelles.

Dans les mégadonnées, la relation homme-données est asymétrique, basée sur le contrôle des données. Le « droit à l'oubli », adopté au niveau de l'UE, est l'un des éléments fondamentaux du contrôle d'un individu sur ses données personnelles.

Transparence

Les algorithmes utilisés dans les mégadonnées peuvent conduire à des biais qui affectent systématiquement les droits de l'individu. Par conséquent, la conception de l'algorithme doit être transparente et inclusive.

La gouvernance anticipative implique une analyse prédictive basée sur les mégadonnées pour évaluer les comportements potentiels, avec des implications éthiques qui peuvent encourager les préjugés et la discrimination.

Une personne qui accepte l'inclusion de ses données personnelles dans les mégadonnées a le droit de savoir pourquoi les données sont collectées, comment elles seront utilisées, combien de temps elles seront stockées et comment elles pourront être modifiées.

Confiance

La confiance dans les systèmes des mégadonnées est liée à l'interdépendance avec la confidentialité et la sensibilisation. Jusqu'à présent, la confiance a été considérée d'un point de vue strictement technologique. On espère que des architectures matérielles et logicielles seront développées qui pourraient accroître la confiance entre les êtres humains et les objets, et donc une plus grande acceptation de l'utilisation des données personnelles.

Propriété

Une question fondamentale dans l'éthique de la recherche sur les mégadonnées est de savoir à qui appartiennent les données ? Cela implique le sujet des droits et obligations de propriété. En droit européen, le RGPD indique que les gens détiennent leurs propres données personnelles.

La somme des données personnelles d'un individu forme une identité numérique.

La protection des droits moraux (le droit d'être identifié comme source de données et de les contrôler) d'un individu repose sur l'opinion que les données personnelles sont une expression

directe de sa personnalité et ne peuvent être transférées qu'à une autre personne, éventuellement, par succession au décès de l'individu.

La propriété implique l'exclusivité, c'est-à-dire la restriction implicite d'autrui concernant l'accès à la propriété. Une propriété efficace des données personnelles implique la portabilité, la possibilité d'utiliser des alternatives sans perdre de données. La normalisation aiderait également à nettoyer les données personnelles.

En fait, à l'heure actuelle, les données sont détenues par le propriétaire des capteurs, celui qui effectue l'enregistrement ou l'entité propriétaire du capteur.

Dans l'UE, la possibilité que les données des citoyens de l'UE soient stockées en dehors de ce que l'on appelle l'« Euro cloud » a été progressivement réduite, mais le problème des données déjà stockées et traitées ailleurs n'a pas été résolu et « ne résout pas le dilemme éthique de la façon dont la propriété des données est définie philosophiquement, avant de passer à une approche plus juridique et politique. » (European Economic and Social Committee 2017)

Surveillance et sécurité

De plus en plus de sources de données sont disponibles à l'aide de technologies avancées telles que la vidéosurveillance, le GPS, les appareils mobiles, les cartes de crédit, les distributeurs automatiques de billets. De plus, la surveillance active est une méthode de collecte de données, mais en même temps limitant les libertés des citoyens. Une telle surveillance permanente détermine l'augmentation du stress des personnes et crée leur tendance à se comporter d'une certaine manière conforme aux normes attendues.

Identité numérique

L'identité numérique a l'avantage d'un accès rapide au contenu en ligne et aux services associés. L'utilisation de l'identité numérique a le potentiel de générer une discrimination fondée

sur la représentation d'une personne en fonction de ses données en ligne, qui peut souvent ne pas correspondre à la situation réelle, dans un processus appelé « dictature des données » dans lequel « nous ne sommes plus jugés sur la base d'actions mais sur la base de ce qui indique toutes les données nous concernant comme nos actions probables », (Norwegian Data Protection Authority 2013) l'interaction personnelle n'est pas placée dans un plan secondaire.

Réalité ajustée

Toute interaction que nous avons avec Internet implique la possibilité de stocker nos données personnelles. Le traitement et l'analyse de ces données déterminent les résultats personnalisés qui apparaissent ultérieurement sur Internet, à travers les résultats de nos recherches, l'affichage des produits dans les boutiques en ligne, l'affichage des publicités, etc. Cela génère une version plus étroite et plus personnalisée de l'expérience en ligne précédente d'un utilisateur (ce que l'on appelle le « ballon filtre ». (Pariser 2011) Un avantage est que l'utilisateur trouvera rapidement ce qu'il recherche habituellement, mais l'exclusion de certains aspects, perspectives et idées peut conduire à une restriction de la créativité et au développement d'une attitude tolérante par isolement politique et social des autres aspects, par le manque de vues pluralistes. (Crawford, Gray, and Miltner 2014)

De-anonymisation

La désidentification implique la suppression ou la dissimulation d'éléments qui pourraient immédiatement identifier une personne ou une organisation. La législation de différents pays sur la protection des données définit différents traitements pour les données identifiables. L'identifiabilité est de plus en plus considérée comme un continuum et non comme un aspect binaire. Les risques de divulgation augmentent simultanément avec le nombre de variables, les sources de données et la puissance de l'analyse des données. Les risques de divulgation peuvent

être atténués mais non éliminés. La désidentification reste un outil essentiel pour garantir une utilisation sûre des données. (UK Data Service 2017)

Des informations parfaitement anonymes prises séparément peuvent être combinées avec d'autres données pour identifier de manière unique une personne avec différents degrés de certitude. Le profilage peut devenir un outil puissant, suscitant des inquiétudes quant à la mesure dans laquelle l'intrusion dans la vie d'un individu est autorisée, la possibilité d'assurer la sécurité et la supervision.

Inégalité numérique

Les avantages d'une grande taille de données sont évidents, mais certains pensent également que l'accumulation de données à grande échelle présente des risques spécifiques. De ce fait, peu d'entités ont accès, via l'infrastructure et les compétences, aux systèmes des mégadonnées. Dans ce contexte, les coûts et les compétences nécessaires à l'accès conduisent à des inégalités numériques spécifiques traitées par l'éthique.

Confidentialité

Dans les transactions de données, il est très important de garantir la confidentialité :

> « Nul ne sera l'objet d'immixtions arbitraires dans sa vie privée, sa famille, son domicile ou sa correspondance, ni d'atteintes à son honneur et à sa réputation. Chacun a droit à la protection de la loi contre de telles ingérences ou attaques. » - *Déclaration des Droits de l'homme des Nations Unies*, Article 12.

Dans de nombreux pays, la surveillance publique des données par le gouvernement pour observer les citoyens nécessite une autorisation explicite par le biais d'un processus judiciaire approprié. La confidentialité ne consiste pas à garder des secrets, mais à choisir, à respecter les droits de l'homme et la liberté.

Souvent, la confidentialité est considérée à tort comme un choix binaire entre l'isolement et le progrès scientifique. La protection de l'identité dans les données est technologiquement possible, par exemple en utilisant un cryptage homomorphe et une conception algorithmique.

La confidentialité en tant que limitation de l'utilisation des données peut également être considérée comme contraire à l'éthique, (Kostkova et al. 2016) en particulier dans les soins de santé, mais il convient de garder à l'esprit qu'il est possible d'extraire la valeur des données sans compromettre la confidentialité.

La confidentialité est reconnue comme un droit de l'homme par de nombreuses réglementations nationales et internationales. La confidentialité dans la recherche est obtenue grâce à une combinaison d’approches : limiter les données collectées, les anonymiser ; et réglementer l'accès aux données. Dans le cas de la recherche des mégadonnées, des problèmes spécifiques se posent : l'ambiguïté entre les termes « vie privée » et « confidentialité »; la déclaration des espaces sociaux publics ou privés; l'ignorance des risques de confidentialité par les utilisateurs; la distinction floue entre les utilisations publiques et privées. Il existe actuellement des différends quant à savoir si la science des données doit être classée comme une recherche sur des sujets humains, et donc non soumis aux règles habituelles de confidentialité.

6. Recherche des mégadonnées

A travers les nouveaux concepts de « dommages algorithmiques », « analyse prédictive », etc., les algorithmes actuellement utilisés dans les opérations avec les mégadonnées dépassent la vision traditionnelle de la confidentialité. Selon le Conseil national pour la science et la technologie des États-Unis,

> « Les « algorithmes analytiques » sont des algorithmes de priorisation, de classification, de filtrage et de prédiction. Leur utilisation peut créer des problèmes de confidentialité lorsque les informations utilisées par les algorithmes sont inadéquates ou inexactes, lorsque des décisions incorrectes se produisent, lorsqu'il n'existe aucun moyen d'appel raisonnable,

> lorsque l'autonomie d'un individu est directement liée au résultat algorithmique ou lors de l'utilisation d'algorithmes prédictifs encourage d'autres atteintes à la vie privée. » (NSTC (National Science and Technology Council) 2016, 18)

La recherche sur les mégadonnées est ce que l'éthicien James Moor qualifierait de « marché conceptuel » en raison de « l'incapacité de conceptualiser correctement les valeurs éthiques et les dilemmes du jeu dans un nouveau contexte technologique ». (Buchanan and Zimmer 2018) Dans cette situation, la confidentialité est assurée par une combinaison de différentes tactiques et pratiques (environnements contrôlés ou anonymes, limitation des informations personnelles, anonymisation des données, restrictions d'accès, sécurité des données, etc.). En général, tous les concepts associés deviennent confus dans le cas des mégadonnées. Ainsi, les publications sociales sont considérées comme publiques sur les réseaux sociaux en cas de paramétrage approprié. Mais les réseaux sociaux sont des environnements complexes d'interactions sociotechniques où les utilisateurs ne comprennent pas toujours la fonctionnalité des paramètres et des conditions d'utilisation. Ainsi, il existe une incertitude quant aux intentions et aux attentes des utilisateurs, et ces lacunes conceptuelles dans le contexte de la recherche sur les mégadonnées conduisent à des incertitudes quant à la nécessité d'un consentement informé.

Conclusions

Les études de données critiques dans les mégadonnées reflètent des pratiques, cultures, politiques et économies spécifiques. (Dalton, Taylor, and Thatcher 2016) Les problèmes peuvent aller de l'intimité et de l'autonomie des individus à l'éthique de la science des données et des changements institutionnels dus à la recherche sur les mégadonnées. Il s'ensuit la nécessité d'analyser les pratiques des mégadonnées conscientes des relations de pouvoir, des préjugés et des inégalités.

Une définition qui limiterait la recherche critique au domaine de la théorie normative et critique serait contre-productive.

Les principes communs des études de données critiques mettent en évidence les interdépendances entre les technologies émergentes et les acteurs (humains) dans des sociétés de plus en plus présentées. Les mégadonnées sont également le produit des conditions socio-techniques contemporaines, car elles produisent de telles conditions. (Richterich, 2018)

Le domaine des études scientifiques et technologiques (EST) a une relation assez ambiguë avec les évaluations normatives de la technologie.

Dans EST, certaines composantes sont davantage concernées par les approches descriptives que normatives.

Contrairement à l'idéal d'EST commun d'un relativisme « sans valeur », (Pels 1996, 277) Pels appelle à la reconnaissance de la « troisièmes positions » dans les évaluations de la production de connaissances scientifiques qui « [...] ne sont pas en dehors du domaine de controverse étudié, mais y sont inclus et impliqués. ... Ils ne sont ni exempts de valeur ni manquants, mais sont localisés partiellement et engagés au sens politique et du savoir. »

Un problème majeur dans les mégadonnées est que les micro-processus empiriques qui sous-tendent l'apparence de leurs caractéristiques de réseau typiques ne sont pas bien connus. (Snijders, Matzat, and Reips 2012) Les mégadonnées devraient toujours être contextualisées dans leurs contextes sociaux, économiques et politiques. (Graham 2012)

Les défenseurs de la vie privée sont préoccupés par la menace à la vie privée en raison de l'augmentation du volume de stockage et de l'intégration des informations personnellement identifiables. À cet égard, il existe différentes recommandations politiques pour respecter la pratique et la vie privée. (Ohm 2012) L'utilisation abusive des mégadonnées par les médias, les

entreprises et même le gouvernement, a entraîné une perte de confiance dans les institutions sociales. Afin de protéger les libertés individuelles, Nayef Al-Rodhan estime qu'un nouveau type de contrat social est nécessaire, avec une surveillance et une réglementation plus stricte des mégadonnées. (Al-Rodhan 2018)

Les expériences scientifiques ont tendance à analyser les données à l'aide de clusters spécialisées et d'ordinateurs hautes performances plutôt que dans le cloud, ce qui différencie culturellement et technologiquement le reste de la société.

L'utilisation des mégadonnées, en raison de la manipulation de grandes quantités de données, a conduit à négliger les principes de la science, tels que le choix d'échantillons représentatifs, provoquant des biais dans l'analyse des résultats. Cette analyse est souvent superficielle par rapport à l'analyse d'ensembles de données plus petits. (Piatetsky 2014) Certaines sources de données, comme Twitter, ne sont pas représentatives de la population totale. Ioannidis a fait valoir qu'en utilisant les mégadonnées, « la plupart des résultats de recherche publiés sont faux » (Ioannidis 2005) car la probabilité qu'un résultat « significatif » soit faux augmente rapidement avec le volume de données, mais seuls les résultats positifs sont publiés.

En utilisant les mégadonnées, UK Data Service met en évidence plusieurs problèmes éthiques spécifiques : (UK Data Service 2017)

- Des alternatives au consentement individuel informé, telles que le « consentement social », sont apparues et sont plus permissives.
- La nécessité de respecter la source des données et, en général, « l'intégrité contextuelle » dans le cas de la réutilisation des données a augmenté.
- L'éthique de la recherche est principalement basée sur l'idée que l'entité recherchée est une personne individuelle, il serait donc possible de se désidentifier pour la protection. Dans le

cas de la prise en compte d'un groupe dans son ensemble, la protection sociale diminue. Dans ce cas, il a été proposé que les données soient considérées comme des « avantages publics » ou « d'intérêt public », mais cela ne résout pas la responsabilité des utilisateurs des données.

Matthew Zook et al. propose « dix règles simples » pour l'utilisation du Big Data dans la recherche. (Zook et al. 2017) Les cinq premières règles concernent la façon de réduire les risques de damage résultant des pratiques de recherche, et les autres règles se réfèrent aux meilleures pratiques.

1. *Les données sont humaines et peuvent nuire* : la plupart des données représentent ou influencent les gens. Commencez avec l'hypothèse que les données sont personnelles (jusqu'à preuve du contraire) et guidez votre analyse sur cette base.
2. *La confidentialité est plus qu'une valeur binaire* : la confidentialité dépend de la nature des données, du contexte dans lequel elles ont été créées et obtenues, ainsi que des attentes et des normes des personnes concernées. Il s'étend aux groupes. Contextualisez les données pour anticiper les atteintes à la confidentialité et minimiser les dommages.
3. *Éviter de réidentifier vos données* : Souvent, les données anonymes ne réussissent pas. Les données considérées comme anonymes sont combinées avec d'autres variables pouvant conduire à une nouvelle identification. Identifier les vecteurs possibles de réidentification et les minimiser dans les résultats publiés.
4. *Pratiquer l'échange éthique de données* : Pour certains projets, tels que la génétique, le partage de données est une nécessité sociale, mais le consentement éclairé et le droit de rétractation restent valables. Partagez les données conformément aux protocoles de

recherche, mais prenez en compte les dommages potentiels générés par les données collectées de manière informelle.

5. *Tenir compte des forces et des limites de vos données* : plus grand ne signifie pas automatiquement meilleur : les ensembles de données doivent être ancrés dans leur contexte, y compris en tenant compte des conflits d'intérêts. Dans l'acquisition de données, il est important de comprendre la source des données et de se conformer à la réglementation. Dans des environnements mal réglementés, des règles éthiques peuvent être utilisées. Les chercheurs doivent être sensibles aux multiples significations potentielles des données. Documentez l'origine et l'évolution des données.
6. *Débattre des choix éthiques difficiles* : le manque de solutions et de protocoles clairs doit être évité. Ces débats peuvent produire des évaluations par les pairs très utiles. Les services de consultation peuvent être utilisés dans le domaine de l'éthique de la recherche dans les universités. Impliquez vos collègues et étudiants dans une pratique éthique pour une recherche des mégadonnées à grande échelle.
7. *Élaborer un code de conduite pour votre organisation, communauté de recherche ou industrie* : la « fausse éthique », ainsi que la falsification de données ou de résultats, sont inacceptables. Il est nécessaire d'élaborer des codes de conduite qui peuvent fournir des orientations dans l'évaluation mutuelle des publications et dans l'examen des financements. Établir des codes de conduite éthique appropriés, avec des représentants des communautés affectées.
8. *Concevoir vos données et vos systèmes d’audit :* l'audit fournit un mécanisme de vérification du travail, améliorant la compréhension et la reproductibilité. Planifier et lancer des audits des pratiques des mégadonnées.

9. *Impliquez-vous avec des conséquences moindres dans les pratiques et l'analyse des données* : il est important pour les chercheurs de penser au-delà des valeurs traditionnelles. Les fournisseurs peuvent être tenus de stocker dans le cloud et les centres de traitement des données peuvent passer à des sources d'énergie durables et renouvelables. La réalisation de recherches à grande échelle a des effets au niveau de la société.
10. *Sacher quand enfreindre ces règles* : vous devez savoir à quoi vous attendre lorsque vous vous éloignez de ces règles, par exemple en cas de catastrophe naturelle ou d'urgence. La recherche responsable des mégadonnées dépend de plusieurs listes de contrôle.

Quelles que soient les normes éthiques ou juridiques, les scientifiques doivent être rigoureux dans l'utilisation des techniques et des méthodologies, et très prudents dans les questions éthiques. L'idée que « les données sont déjà publiques » (Zimmer 2016) est une simplification injustifiée. Les données ne sont pas abstraites, ce sont en fait de vraies personnes.

La recherche responsable sur les mégadonnées ne vise pas à restreindre la recherche, mais à assurer la confiance, l'équité et à maximiser les aspects positifs tout en réduisant les dommages. Les mégadonnées offrent des opportunités fantastiques pour une meilleure compréhension de la société et du monde, mais la responsabilité éthique dans les choix, les pratiques et les actions de recherche doit également être prise en compte.

Bibliographie

Allen, Marshall. 2018. “Health Insurers Are Vacuuming Up Details About You — And It Could Raise Your Rates.” Text/html. ProPublica. July 17, 2018. https://www.propublica.org/article/health-insurers-are-vacuuming-up-details-about-you-and-it-could-raise-your-rates.

Al-Rodhan, Nayef. 2018. “The Social Contract 2.0: Big Data and the Need to Guarantee Privacy.” OpenMind. June 11, 2018. https://www.bbvaopenmind.com/en/humanities/beliefs/the-social-contract-2-0-big-data-and-the-need-to-guarantee-privacy-and-civil-liberties/.

Baert, Patrick, and Bryan Turner. 2004. “New Pragmatism and Old Europe: Introduction to the Debate between Pragmatist Philosophy and European Social and Political Theory.” *European Journal of Social Theory* 7 (3): 267–74. https://doi.org/10.1177/1368431004044193.

Brumfiel, Geoff. 2011. “High-Energy Physics: Down the Petabyte Highway.” *Nature* 469 (7330): 282–83. https://doi.org/10.1038/469282a.

Buchanan, Elizabeth A., and Michael Zimmer. 2018. “Internet Research Ethics.” In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Winter 2018. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/win2018/entries/ethics-internet-research/.

Collins, Tom. 2016. “Security Fears Sparked over Wearable Technology.” Mail Online. December 19, 2016. http://www.dailymail.co.uk/~/article-4049154/index.html.

Constine, Josh. 2017. “Facebook Now Has 2 Billion Monthly Users… and Responsibility.” *TechCrunch* (blog). 2017. http://social.techcrunch.com/2017/06/27/facebook-2-billion-users/.

Crawford, Kate, Mary L. Gray, and Kate Miltner. 2014. “Big Data| Critiquing Big Data: Politics, Ethics, Epistemology | Special Section Introduction.” *International Journal of Communication* 8 (0): 10. https://ijoc.org/index.php/ijoc/article/view/2167.

Cumbley, Richard, and Peter Church. 2013. “Is Big Data Creepy.” In . https://doi.org/10.1016/j.clsr.2013.07.007.

Dalton, Craig M., Linnet Taylor, and Jim Thatcher. 2016. “Critical Data Studies: A Dialog on Data and Space.” In . https://doi.org/10.1177/2053951716648346.

Dean, Jeffrey, and Sanjay Ghemawat. 2004. “MapReduce: Simplified Data Processing on Large Clusters.” http://static.googleusercontent.com/media/research.google.com/en//archive/mapreduce-osdi04.pdf.

Dedić, Nedim, and Clare Stanier. 2017. “Towards Differentiating Business Intelligence, Big Data, Data Analytics and Knowledge Discovery.” In *Innovations in Enterprise Information Systems Management and Engineering*, edited by Felix Piazolo, Verena Geist, Lars Brehm, and Rainer Schmidt, 114–22. Lecture Notes in Business Information Processing. Springer International Publishing.

DG Connect. 2015. “Study on Personal Data Stores Conducted at the Cambridge University Judge Business School.” Text. Digital Single Market - European Commission. August 7, 2015. https://ec.europa.eu/digital-single-market/en/news/study-personal-data-stores-conducted-cambridge-university-judge-business-school.

European Commission. 2019. “Horizon 2020.” Text. Horizon 2020 - European Commission. 2019. https://ec.europa.eu/programmes/horizon2020/en.

European Economic and Social Committee. 2017. "The Ethics of Big Data: Balancing Economic Benefits and Ethical Questions of Big Data in the EU Policy Context." European Economic and Social Committee. February 22, 2017. https://www.eesc.europa.eu/en/our-work/publications-other-work/publications/ethics-big-data.

European Parliament. 2016. *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the Protection of Natural Persons with Regard to the Processing of Personal Data and on the Free Movement of Such Data, and Repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA Relevance)*. *OJ L*. Vol. 119. http://data.europa.eu/eli/reg/2016/679/oj/eng.

Everts, Sarah. 2016. "Information Overload." Science History Institute. July 18, 2016. https://www.sciencehistory.org/distillations/magazine/information-overload.

Floridi, Luciano. 2014. *The Fourth Revolution: How the Infosphere Is Reshaping Human Reality*. OUP Oxford.

Graham, Mark. 2012. "Big Data and the End of Theory?" *The Guardian*, March 9, 2012, sec. News. https://www.theguardian.com/news/datablog/2012/mar/09/big-data-theory.

Ioannidis, John P. A. 2005. "Why Most Published Research Findings Are False." *PLOS Medicine* 2 (8): e124. https://doi.org/10.1371/journal.pmed.0020124.

Keulartz, Jozef, Maartje Schermer, Michiel Korthals, and Tsjalling Swierstra. 2004. "Ethics in Technological Culture: A Programmatic Proposal for a Pragmatist Approach." *Science, Technology, & Human Values* 29 (1): 3–29. https://doi.org/10.1177/0162243903259188.

Kostkova, Patty, Helen Brewer, Simon de Lusignan, Edward Fottrell, Ben Goldacre, Graham Hart, Phil Koczan, et al. 2016. "Who Owns the Data? Open Data for Healthcare." *Frontiers in Public Health* 4. https://doi.org/10.3389/fpubh.2016.00007.

Kvochko, Elena. 2012. "Four Ways to Talk About Big Data." Text. Information and Communications for Development. December 4, 2012. http://blogs.worldbank.org/ic4d/four-ways-to-talk-about-big-data.

Manyika, James, Michael Chui, Jaques Bughin, and Brad Brown. 2011. "Big Data: The next Frontier for Innovation, Competition, and Productivity." 2011. https://www.mckinsey.com/business-functions/digital-mckinsey/our-insights/big-data-the-next-frontier-for-innovation.

Mayer-Schönberger, Viktor, and Kenneth Cukier. 2014. *Big Data: A Revolution That Will Transform How We Live, Work, and Think*. Reprint edition. Boston: Eamon Dolan/Mariner Books.

Mearian, Lucas. 2015. "Insurance Company Now Offers Discounts -- If You Let It Track Your Fitbit." Computerworld. April 17, 2015. https://www.computerworld.com/article/2911594/insurance-company-now-offers-discounts-if-you-let-it-track-your-fitbit.html.

MIKE2.0. 2019. "Big Data Solution Offering - MIKE2.0, the Open Source Methodology for Information Development." 2019. http://mike2.openmethodology.org/wiki/Big_Data_Solution_Offering.

MORO, Veronica. 2016. "Meeting the Challenges of Big Data." Text. European Data Protection Supervisor - European Data Protection Supervisor. November 16, 2016. https://edps.europa.eu/data-protection/our-work/publications/opinions/meeting-challenges-big-data_en.

Nature. 2008. "Community Cleverness Required." *Nature* 455 (7209): 1. https://doi.org/10.1038/455001a.

Noorman, Merel. 2012. "Computing and Moral Responsibility." *Stanford Encyclopedia of Philosophy*.
Norwegian Data Protection Authority. 2013. "Big Data – Privacy Principles under Pressure." https://www.datatilsynet.no/globalassets/global/english/big-data-engelsk-web.pdf.
NSTC (National Science and Technology Council). 2016. "National Privacy Research Strategy." https://obamawhitehouse.archives.gov/sites/default/files/nprs_nstc_review_final.pdf.
OECD. 2013. "New Data for Understanding the Human Condition: International Perspectives." http://www.oecd.org/sti/inno/new-data-for-understanding-the-human-condition.pdf.
———. 2016. "Research Ethics and New Forms of Data for Social and Economic Research," November. https://doi.org/10.1787/5jln7vnpxs32-en.
Ohm, Paul. 2012. "Don't Build a Database of Ruin." *Harvard Business Review*, August 23, 2012. https://hbr.org/2012/08/dont-build-a-database-of-ruin.
Pariser, Eli. 2011. *The Filter Bubble: What The Internet Is Hiding From You*. Penguin Books Limited.
Pels, Dick. 1996. "The Politics of Symmetry." *Social Studies of Science* 26 (2): 277–304. https://doi.org/10.1177/030631296026002004.
Piatetsky, Gregory. 2014. "Interview: Michael Berthold, KNIME Founder, on Research, Creativity, Big Data, and Privacy, Part 2." 2014. https://www.kdnuggets.com/2014/08/interview-michael-berthold-knime-research-big-data-privacy-part2.html, https://www.kdnuggets.com/2014/08/interview-michael-berthold-knime-research-big-data-privacy-part2.html.
Reichman, O. J., Matthew B. Jones, and Mark P. Schildhauer. 2011. "Challenges and Opportunities of Open Data in Ecology." *Science* 331 (February): 703. https://doi.org/10.1126/science.1197962.
Richterich, A. 2018. "The Big Data Agenda: Data Ethics and Critical Data Studies." https://doi.org/10.16997/book14.b.
Rip, Arie. 2013. "Pervasive Normativity and Emerging Technologies." In *Ethics on the Laboratory Floor*, edited by Simone van der Burg and Tsjalling Swierstra, 191–212. London: Palgrave Macmillan UK. https://doi.org/10.1057/9781137002938_11.
Snijders, Chris, Uwe Matzat, and Ulf-Dietrich Reips. 2012. "'Big Data': Big Gaps of Knowledge in the Field of Internet Science." http://www.ijis.net/ijis7_1/ijis7_1_editorial.pdf.
Sullivan, Danny. 2015. "Google Still Doing At Least 1 Trillion Searches Per Year." Search Engine Land. January 16, 2015. https://searchengineland.com/google-1-trillion-searches-per-year-212940.
Tanner, Adam. 2014. "Different Customers, Different Prices, Thanks To Big Data." Forbes. 2014. https://www.forbes.com/sites/adamtanner/2014/03/26/different-customers-different-prices-thanks-to-big-data/.
UK Data Service. 2017. "Big Data and Data Sharing: Ethical Issues." https://www.ukdataservice.ac.uk/media/604711/big-data-and-data-sharing_ethical-issues.pdf.
Zimmer, Michael. 2016. "OkCupid Study Reveals the Perils of Big-Data Science." *Wired*, May 14, 2016. https://www.wired.com/2016/05/okcupid-study-reveals-perils-big-data-science/.
Zook, Matthew, Solon Barocas, Danah Boyd, Kate Crawford, Emily Keller, Seeta Peña Gangadharan, Alyssa Goodman, et al. 2017. "Ten Simple Rules for Responsible Big Data

Research." *PLOS Computational Biology* 13 (3): e1005399. https://doi.org/10.1371/journal.pcbi.1005399.